CJBB75 – 1 (G1310.00-11.30)

K. Osolsobě

Výuka: od 28. 2. 2022 každých 14 dní kontaktní, každých 14 dní úkol (viz harmonogram)

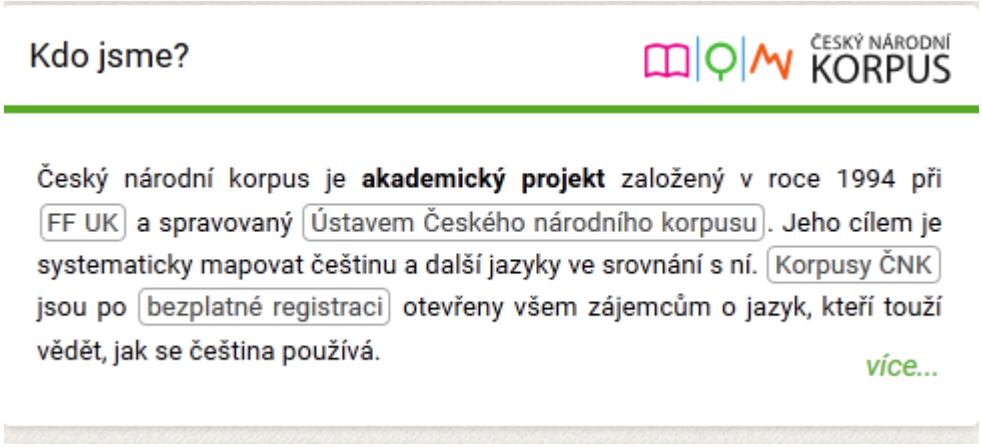
Podmínky ukončení: Průběžné plnění úkolů (5 odevzdaných referátů, závěrečný test).

Náplň dnešní hodiny 21. 2. 2022

Co je to korpus?

- Soubor textů
- elektronicky uložených a přístupných (korpusové manažery – programy, skrze něž lze ke korpusům přistupovat)
- má stanovený obsah (složený z textů záměrně vybraných dle zveřejněných kritérií)
- má stanovený rozsah/velikost (lze na něm pracovat s frekvenčními/kvantitativně měřitelnými údaji)
- obsahuje standardní anotace (metadata – údaje o každém textu a lingvistické interpretace, anotace jazykových jednotek – vnitřní anotace)

Registrace uživatele pro práci s ČNK (http://ucnk.ff.cuni.cz/)



Korpusový manažer

Základy práce s korpusem přes Kontext

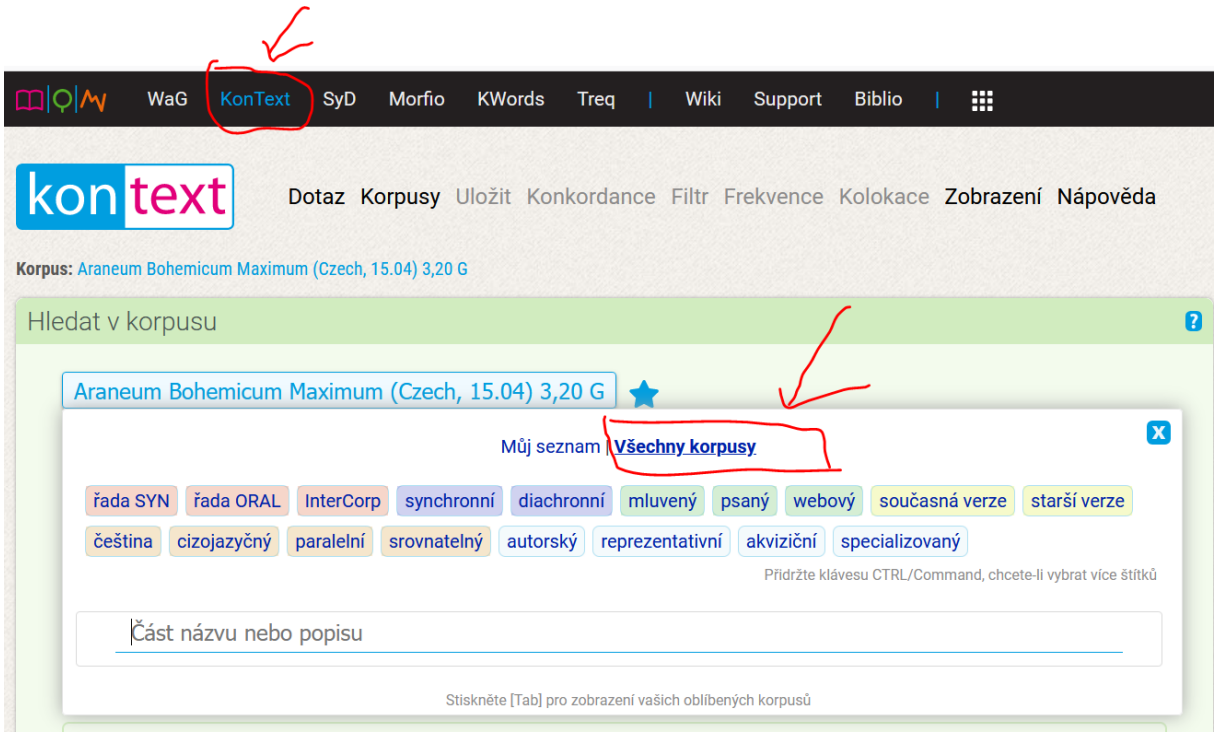
KonText

Dotaz

Výběr korpusu

https://wiki.korpus.cz/doku.php/cnk:uvod

V rozhraní KonText



Jaké korpusy jsou k dispozici ?

Časové hledisko (synchronní / diachronní)

Hledisko textů (psané / mluvené, připravené/spontánní)

Hledisko žánru (vyvážené žánrově/ žánrově kompaktní – např. korpusy výhradně publicistické, nebo korpus soukromé korespondence, projekt Korpus českého verše).

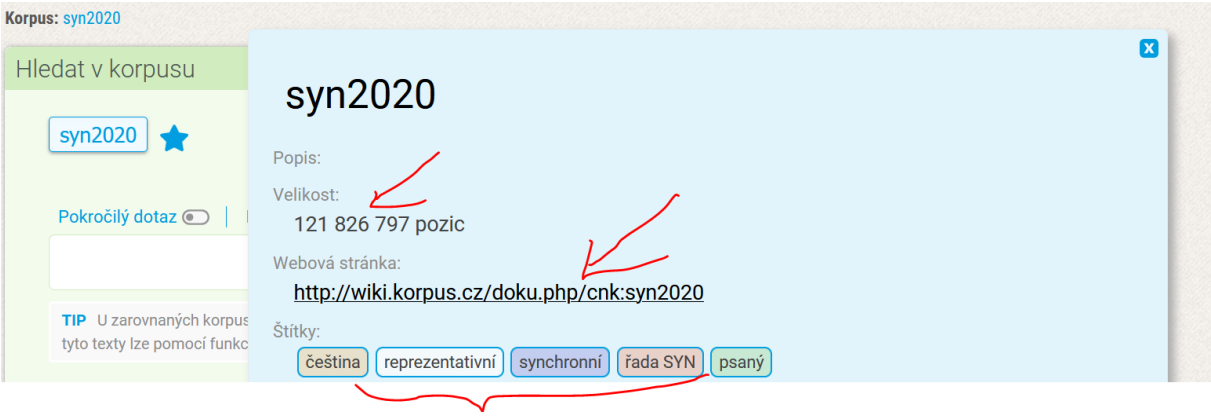
Hledisko autora (autoři jsou rodilí mluvčí/ autoři se učí jazyk, v němž jsou texty vytvořeny jako tzv. druhý jazyk – learner corpora/žákovské korpusy, autorské korpusy založené na díle/korespondenci význačných osobností).

Hledisko jazyka (jednojazyčné – např. čeština/ vícejazyčné, srovnatelné, paralelní).

Vícejazyčné paralelní korpusy – stejné texty – originál+překlad – zarovnaní/alignment = jednotky, které si odpovídají, jsou propojeny / srovnatelné korpusy – různojazyčné i stejného jazyka vybudované stejným způsobem, mající stejné složení).

Jak čteme informace o zvoleném korpusu?

Stručné info.



Citace:

Citační informace:

Korpus jako zdroj dat:

Křen, M. – Cvrček, V. – Henyš, J. – Hnátková, M. – Jelínek, T. – Kocek, J. – Kováříková, D. – Křivan, J. – Milička, J. – Petkevič, V. – Procházka, P. – Skoumalová, H. – Šindlerová, J. – Škrabal, M.: SYN2020: reprezentativní korpus psané češtiny. Ústav Českého národního korpusu FF UK, Praha 2020. Dostupný z WWW: http://www.korpus.cz

Obecné odkazy:

http://www.korpus.cz

Proč je třeba citovat korpusy?

Korpusy ÚČNK vznikly jako výsledek státní podporu GAČR. V korpusech jsou texty, které mnohdy spadají pod autorská práva. ÚČNK poskytl smluvní garance „poskytovatelům textů“.

Jak číst informace o velikosti korpusu:

Termíny: viz http://wiki.korpus.cz/doku.php

http://wiki.korpus.cz/doku.php/pojmy:prehled_pojmu

Tokenizace

Token je nejmenší jednotka textu, většinou se jedná o grafické slovo (tj. řetězec alfabetických znaků oddělený mezerou v textu), resp. o jednu jeho konkrétní realizaci. V některých případech je jedno grafické slovo rozděleno na dvě (např. *mohu -li*), často je také z praktických důvodů (pro snadné vyhledávání) oddělována interpunkce od předcházejícího slova (3 tokeny: *řekl , že*). O jednotlivých tokenech v korpusu se také mluví jako o pozicích.

Všimněme si:

A teď si vemte , že se	ptám	na maličkost , kterou by mi řekl kdejaký úředník .
- Podle policistů z šumperského dopravního inspektorátu se lidé nejčastěji	ptají	na používání bezpečnostních pásů při přepravě dětí . " Je
to , proč se tak doposud nestalo , se nás	ptal	i místostarosta Pavelec , “ vyjádřila se Kelárková s tím
Skalice u České Lípy – Když se ho televizní reportér	ptal	, jaká by měla být jeho vyvolená , odpověděl naprosto
koupit už v pátek . " Lidé se po něm	ptali	. Prodáváme ho s předstihem , protože v neděli máme
, a chce slyšet můj názor . Včera se mě	ptal	třeba na to , jestli nevím , co by se
k nám nastěhovat . První , na co se totiž	ptají	, je , jestli je v Zádveřicích škola . Také
projekty pro děti . Již během roku se mne lidé	ptají	, jaká překvapení zase připravíme . Moc bych chtěla poděkovat
tu nejsou všichni nejlepší , ale na to se historie	neptá	. « Sám říkal , že pokud nezíská medaili ,
pojistných podmínek , protože to je součást jejich školení .	Ptali	jsme se , co by se stalo , kdybychom to
nic jsem okolo sebe nevnímala , protože jsem se znovu	ptala	sama sebe , proč šel zase běhat . Tak jsem
bez onoho jmění neexistovaly . V nadpisu jsem se vás	ptala	dotazem , který na svých webových stránkách pokládají autoři výše
španělsky . „ Ty ses to někde učila ? “	ptal	jsem se . „ Ne , “ odpověděla , „
lidé , kteří jsou v pivovaru na exkurzích , udiveně	ptají	, proč mají téměř prázdné sklady sudového a lahvového piva
. Mimochodem (tento dopis je plný mimochodů) ,	ptal	ses mne z Prahy , zda Chris stále cituje duchy

Pro rodilého mluvčího je zvratné *se* součástí reflexiva tantum *ptát se*. Pro automatickou morfologickou analýzu jde o dva samostatné tokeny.

Pozice

V souvislosti s tím, že každý text, který vstupuje do korpusu, prochází procesem tokenizace, se o jednotkách v korpusu nemluví jako o slovech, ale častěji jako o **pozicích**. Tokenizace se přitom u jednotlivých korpusů může lišit, pozicí se tak v různých korpusech může myslet různě vymezená jednotka.

Lemmatizace a taggování

Lemma je reprezentativní slovníková podoba hesla, při automatickém zpracování jazyka je pak tato podoba v procesu lemmatizace přidělována každé formě v korpusu.

Přístupy k lemmatizaci se mohou v drobnostech lišit, obecně však platí, že

- lemma každého českého substantiva je jeho **nom. sg.** (tvary *lesům, lesy, lesích* mají lemma ***les***)
- u adjektiv je to **nom. sg. mask.pozitiv** (tvary *chytrého, chytrou, chytrejma, nejchytřejší* mají lemma ***chytrý***)
- u zájmen je to **nom. sg. mask.** (tvary *ta, to, ti, tomu* mají lemma ***ten***)
- u sloves je to **infinitiv** (tvary *chodil, chodíš, chodíme* mají lemma ***chodit***)

Lemma jako jednotka vzniká abstrakcí morfologických vlastností slovního tvaru (označovaného jako word nebo forma), představuje tedy množinu forem se stejným kořenem lišící se pouze morfologickými afixy, příp. pravopisnou variantou. V některých koncepcích se pak k lemmatu řadí i vybrané varianty slovotvorné.

Představme si následující dialog, která z variant je podle vás více na místě, A nebo B?

A

- No víš, viděl jsem takovou ***fuškunkuli*** a ona ti měla na hlavě takovou ***kumušinku paškovanou*** a ona si ji ještě ***vygárovala***.
- Co je to ***fuškunkuli*** a ***kumušinku paškovanou***? A co je to ***vygárovala***?

B

- No víš, viděl jsem takovou ***fuškunkuli*** a ona ti měla na hlavě takovou ***kumušinku paškovanou*** a ona si ji ještě ***vygárovala***.
- Co je to ***fuškunkule*** a ***kumušinka paškovaná***? A co je to ***vygárovat***?

Uvědomte si, že lemmatizace je činnost, kterou dnes automaticky provádí řada nástrojů od vyhledávačů na webu přes on-line slovníky. Jde ale o schopnost, kterou nabývá i dítě během akvizice jazyka, kterou má mluvčí, když se dotazuje na neznámé slovo, kterou aplikujeme, když hledáme v cizojazyčném slovníku (např. význam tvaru *went* nenajdeme ve slovníku angličtiny pod ***w***, ale pod ***g***).

Desambiguace

Desambiguace (někdy též disambiguace, z lat. *dis-* vyjadřuje zápor, *ambo* oba, česky zjednoznačnění) je část (většinou automatického) procesu anotace jazykových dat, které vstupují do korpusu.

Zjednoznačněním se většinou myslí odstranění homonymie, čili jednoznačná interpretace slovního tvaru či skupiny slovních tvarů nebo věty na základě kontextu či mimojazykové situace. Desambiguace se obecně týká všech jazykových rovin, nejčastěji se ovšem v korpusech češtiny uplatňuje na rovině morfologické (zahrnující lemmatizaci a přiřazení náležitých morfologických údajů slovnímu tvaru na základě kontextu).

Např. ve větě *Větry vanou od západu.* se při morfologické interpretaci věty nejprve přiřadí morfologickou analýzou tvaru *vanou* dvě lemmata a dvě morfologické interpretace:

1. lemma = *vana*, subst. fem. sg. instr.
2. lemma = *vát*, 3. os. pl. préz,

a poté se při desambiguaci vybere náležitá 2. interpretace.

V následujících větách si všimněte, jak je třeba nejednoznačný tvar ***sil***, který lze interpretovat jako a) genitiv plurálu feminina k lemmatu ***síla***, b) genitiv plurálu neutra k lemmatu ***silo***, c) variantní tvar l-ového příčestí maskulina singuláru slovesa ***sít***.

*a) Podle jeho názorů je internet jednou ze **sil**, která dostala Ameriku na špici*

*b) Z jednoho ze **sil** začala náhle tryskat čpící tekutina a ocelová konstrukce jedné z věží se zhroutila.*

*c) Raná variační fantazie na lidový nápěv **Sil** jsem proso dala oběma protagonistům možnost ukázat jejich virtuozitu.*

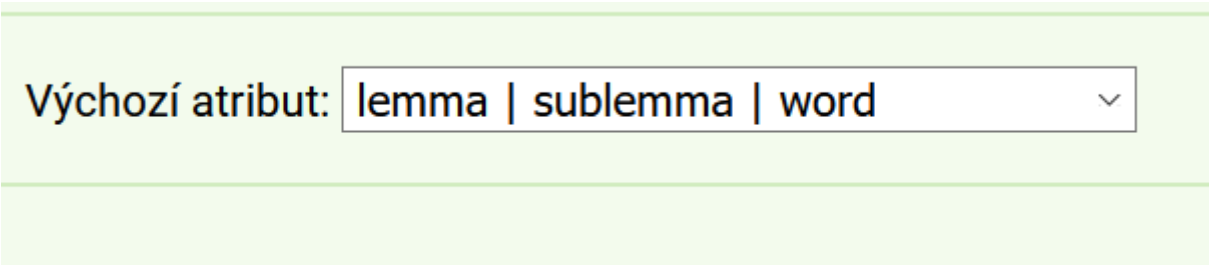
Někdy může být situace dosti složitá:

*Odstupující ministr informatiky Vladimír Mlynář podle serveru iDNES odmítl nabídku premiéra Grosse **stát se** šéfem Českého telekomunikačního úřadu.*

*Potřeboval **stát se** svým zločineckým gangem.*

Jaké přednosti má lemmatizovaný a morfologicky označkovaný korpus?

Možnosti vyhledávání v korpusu:



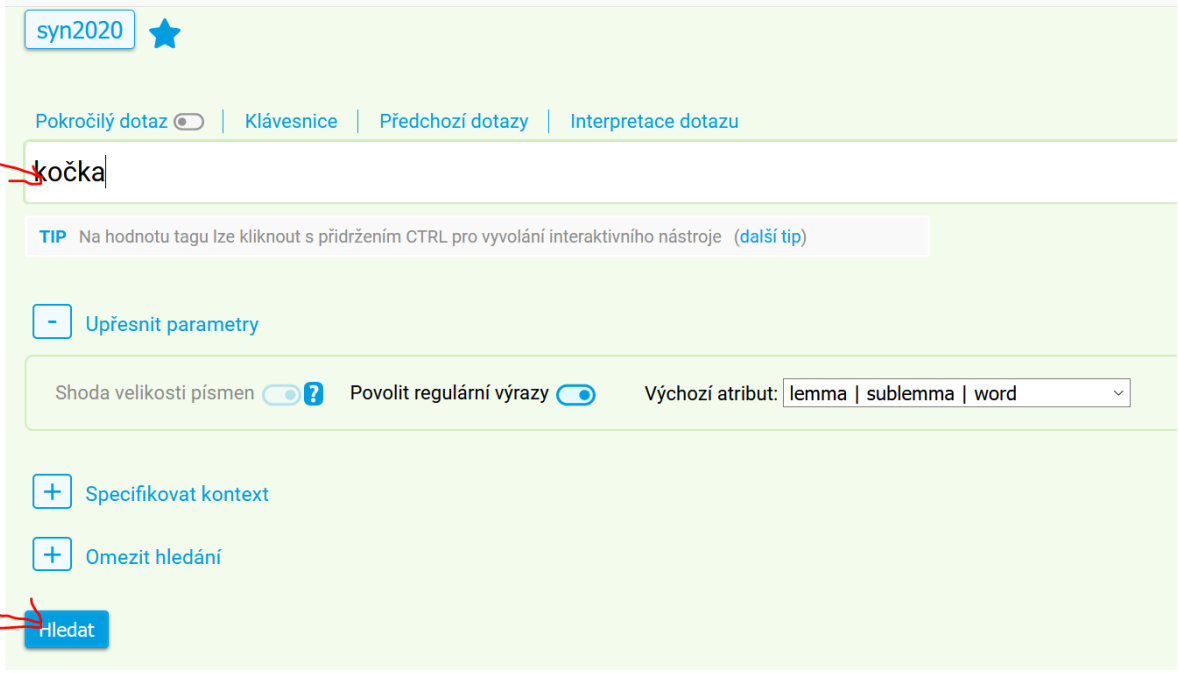
Nabídka výchozího atributu je závislá na konkrétním korpusu, na použité lemmatizaci a značkování.

Regulární výrazy (http://wiki.korpus.cz/doku.php/pojmy:regularni_vyrazy)

Konkordance, KWIC

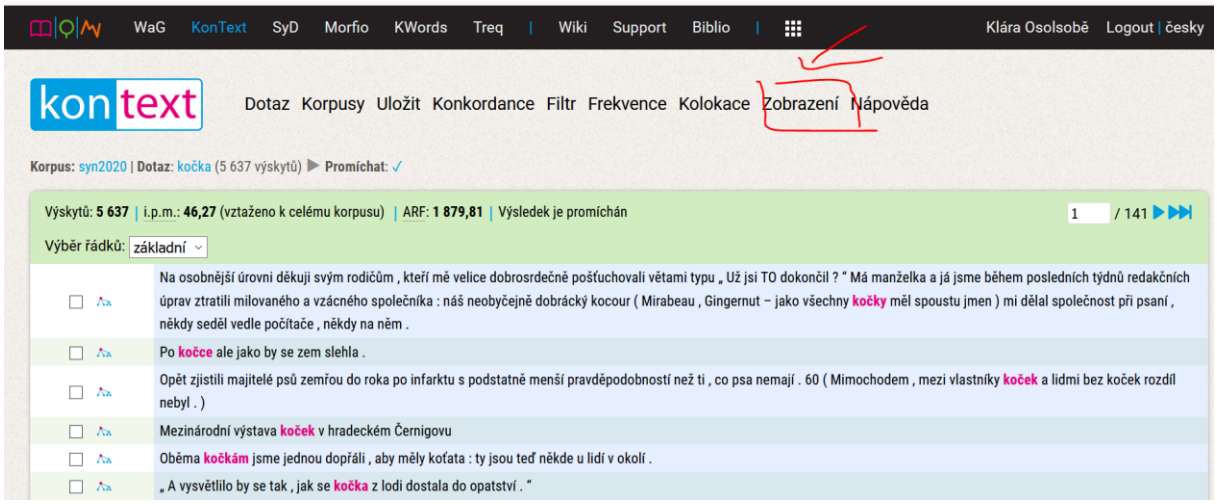
Konkordance představuje všechny doklady (výskyty) hledaného jevu v korpusu spolu s **okolním kontextem**. V praxi se v rámci konkordance rozlišuje **KWIC** (tj. key word in

context), tedy hledané slovo/jev a jeho pravý a levý kontext. Jeden řádek konkordančního seznamu se označuje jako konkordanční řádek.



Zobrazení

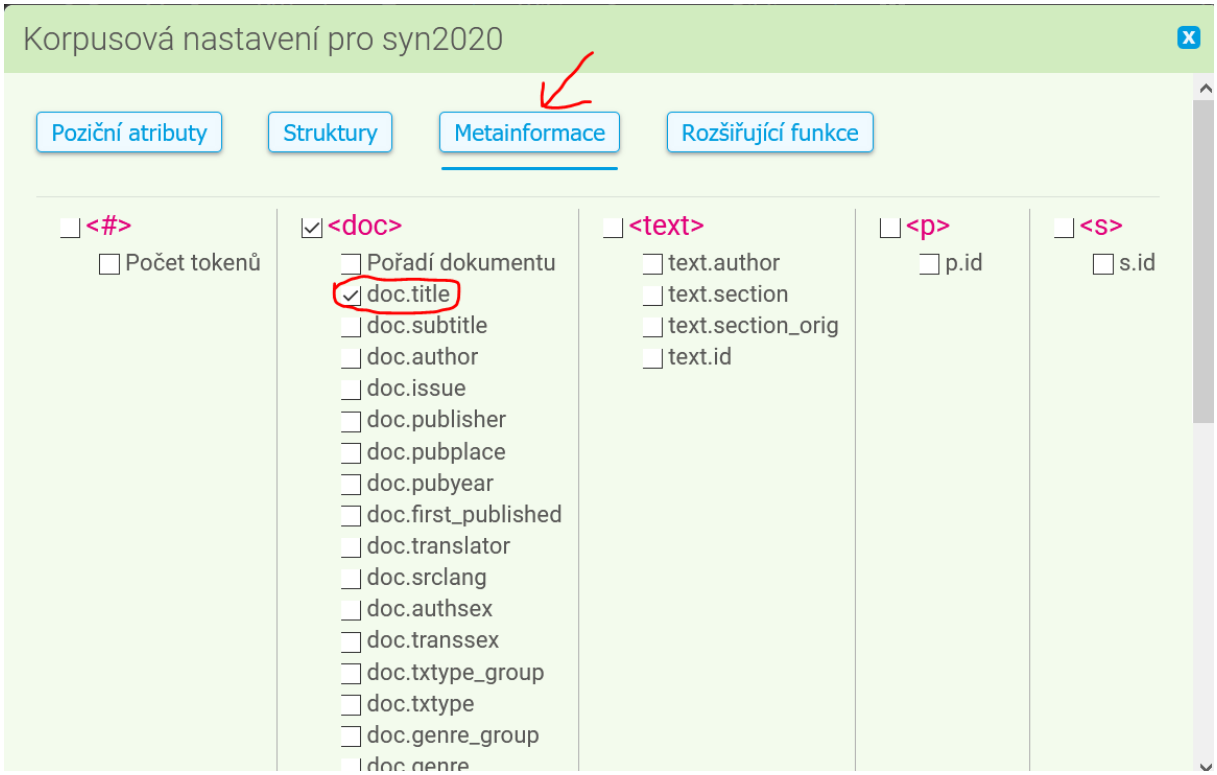
KWIC/Věta



Korpusová nastavení

(Lemma, POS – part of speech)

Metainformace



Kompletní info o zdrojovém textu:

doc.title	Geniální psi	**doc.subtitle**	
doc.author	Hare, Brian - Woods, Vanessa	**doc.issue**	
doc.publisher	Dokořán	**doc.pubplace**	Praha
doc.pubyear	2016	**doc.first_published**	2016
doc.translator	Houserová, Jana - Houser, Pavel	**doc.srclang**	en: angličtina
doc.authsex	X: neuvedeno	**doc.transsex**	X: neuvedeno
doc.txtype_group	NFC: oborová literatura	**doc.txtype**	POP: populárně naučná literatura
doc.genre_group	NAT: přírodní vědy	**doc.genre**	AGR: zemědělství, chovatelství
doc.medium	B: kniha	**doc.periodicity**	NP: neperiodická publikace
doc.audience	GEN: obecné publikum	**doc.isbnissn**	978-80-7363-775-0
doc.biblio	Hare, Brian - Woods, Vanessa (2016): Geniální psi. Překlad: Houserová, Jana - Houser, Pavel. Praha: Dokořán.	**doc.id**	hare_genialnips
text.author		**text.section**	
text.section_orig		**text.id**	hare_genialnips:1
p.id	hare_genialnips:1:1074	**s.id**	hare_genialnips:1:1074:4

Cvičení:

Najděte v korpusu SYN2020 slovní tvar (atribut **word**) ***jedle***.

Jak je tvar interpretován lemmatem a tagem?

Najděte v korpusu SYN2020 slovní spojení ***příliš jedle, vypadá jedle***.

Jak je tvar interpretován lemmatem a tagem?

přehnané , ale fialová zelenina se zrcadlovým leskem opravdu nevypadá **příliš/příliš/Db**---------- **jedle/jedle/NNFS1-----A----** , navíc syrový lilek (především mladý a sklizený před

kontext Dotaz Korpusy Uložit Konkordance Filtr Frekvence Kolokace Zobrazení Nápověda

Korpus: syn v8 | **Dotaz:** vypad.*, jedle (8 výskytů) ▶ **Promíchat:** ✓

Výskytů: 8 | i.p.m.: **0** (vztaženo k celému korpusu) | ARF: **5,1** | Výsledek je promíchán 1 / 1

Výběr řádků: základní

☐	vajíčka . Zastavuji se u pultu s uzeninami . Čabajka	**vypadá/vypadat/VB-S---3P-AA---I jedle/jedle/NNFS2-----A-----**	. A ty párečky . Dívám se do peněženky .
☐	první jarní květy . A teď mi řekněte , jak	**vypadá/vypadat/VB-S---3P-AA---I jedle/jedle/NNFS1-----A-----**	Děti z hradecké mateřské školy Klíček se vč
☐	, slouží často jako dekorace . Mangostan Na první pohled	**nevypadá/vypadat/VB-S---3P-NA---I jedle/jedle/NNFS1-----A-----**	: slupka je tmavá , až do černa . Když
☐	do původní polohy , " přemítá . Na konci roku	**vypadala/vypadat/VpFS---3R-AA---I jedle/jedle/NNFS2-----A-----**	takhle … … a teď stojí znovu na svém místě
☐	dětech , které ve zdevastované půdě hledají cokoli , co	**vypadá/vypadat/VB-S---3P-AA---I jedle/jedle/NNFS1-----A-----**	" . Potravinové příděly pro řadové obyvatele
☐	Je libo návod pro inspiraci ? Ve společné skupině dobře	**vypadá/vypadat/VB-S---3P-AA---I jedle/jedle/NNFS1-----A-----**	ojíněná , borovice wintergold , která na zim
☐	být podle doktorky Lenderové jakákoli zelená dřevina . Nejlépe samozřejmě	**vypadá/vypadat/VB-S---3P-AA---I jedle/jedle/NNFS1-----A-----**	, ale svůj úkol důstojně plní i smrk nebo bor
☐	často bývají napadeny plísněmi . A tak , i když	**vypadají/vypadat/VB-P---3P-AA---I jedle/jedle/NNFP1-----A-----**	, mohou být zrádné . Mění se u nich i

Úkol na příště:

Prostudovat www stránky ÚČNK

Umět odpovědět na otázky:

1. Co je to korpusu?
2. Co je to Český národní korpus?
3. Jaké typy korpusů máme k dispozici?
4. Co to znamená, když řeknu, že korpus má 100 milionů slov?
5. Jak komunikujeme s korpusem (jak jej můžeme využívat pro lingvistickou práci)?
6. Jak můžeme vyhledat v korpusu výskyt slova, jak se se zobrazí v korpusu výskyt slova a co můžeme se zobrazenými výskyty dále dělat?
7. Jak můžeme vyhledat v korpusu všechna slova, která mají společnou vlastnost, že jsou tvary jednoho základního tvaru?
8. Jak můžeme v korpusu vyhledat všechny tvary na rovině gramatické abstrakce (třeba podstatné jména rodu ženského ve 3. pádě, nebo slovesa v přítomném čase v první osobě)?

A připravit si otázky, na něž byste rádi znali odpověď (souvisí s korpusy!!)